[0008]

[Means for solving problem] The HTML descriptive information extraction display system in the first invention comprises a character string retrieving means for retrieving a keyword character string including a keyword that has been specified by descriptive information described in Hyper Media Description Language, a tag analysis means for analyzing a tag in the descriptive information and detecting a divided tag that has been specified in advance, an information extraction means for extracting a keyword character string from the descriptive information on the basis of the divided tag that has been extracted and forming it into excerpted information described in Hyper Media Description Language, and a display means for displaying the extracted keyword character string.

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－134341

(43)公開日　平成11年(1999)５月21日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３２０Ａ
		15/40	３４０
		15/403	３８０Ｅ

審査請求　有　　請求項の数７　ＯＬ　（全　６　頁）

(21)出願番号　特願平9－292806

(22)出願日　平成９年(1997)10月24日

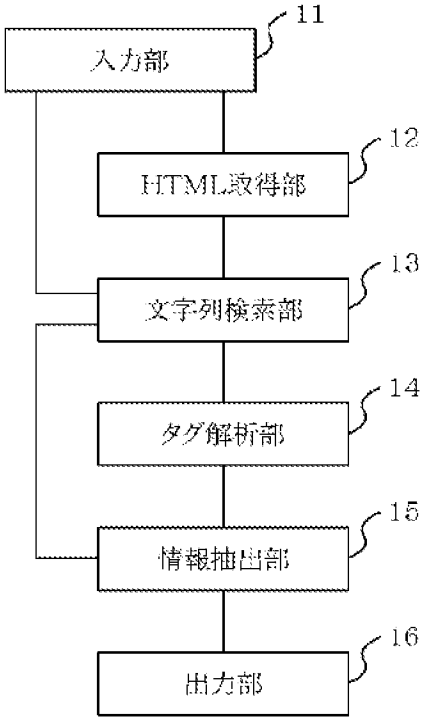
(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　新井　良和
東京都港区芝五丁目７番１号　日本電気株式会社内

(74)代理人　弁理士　京本　直樹　（外２名）

(54)【発明の名称】　ハイパーメデイア記述言語による記述情報の抜粋表示システム

(57)【要約】
【課題】ＨＴＭＬ記述情報の中から必要な部分を自動的に抜粋して一つに纏めて表示することができ、本当に必要な情報の掲載されたページを迅速にかつ、効率的に探索することができるＨＴＭＬ記述情報抜粋表示システムの提供。
【解決手段】入力部１１から入力されたＵＲＬによりＨＴＭＬ取得部１２が取得したＨＴＭＬファイルを文字列検索部１３によりキーワードを含む文字列を検索し、この文字列に含まれている分割タグをタグ解析部１４により見つけてこれによりキーワードを含む文字列を情報抽出部１５により抽出して出力部１６に表示する。



1

【特許請求の範囲】

【請求項１】　ハイパーメデイア記述言語により記述された記述情報から指定されたキーワードを含むキーワード文字列を検索する文字列検索手段と、前記記述情報にあるタグを解析し予め指定された分割タグを検出するタグ解析手段と、前記検出された分割タグに基づいて前記キーワード文字列を前記記述情報から抽出しハイパーメデイア記述言語による抜粋情報に整形する情報抽出手段と、前記抽出されたキーワード文字列を表示する表示手段とを含むことを特徴とするハイパーメデイア記述言語による記述情報の抜粋表示システム。

【請求項２】　文字列検索手段は検索に際しては先ず記述情報からタグを削除してキーワード文字列を検索し検索後に前記削除したタグを復元することを特徴とする請求項１記載のハイパーメデイア記述言語による記述情報の抜粋表示システム。

【請求項３】　情報抽出手段はキーワードに先行する最も近い先行指定分割タグとキーワードに後続する最も近い後続指定分割タグとを検索して前記先行指定分割タグと後続指定分割タグとを含むその間の文字列を抽出することを特徴とする請求項１記載のハイパーメデイア記述言語による記述情報の抜粋表示システム。

【請求項４】　情報抽出手段は抽出したキーワードを含む文字列ブロックが複数あるときには文字列区切りタグにより１つの文字列ブロックにすることを特徴とする請求項３記載のハイパーメデイア記述言語による記述情報の抜粋表示システム。

【請求項５】　情報抽出手段は抽出した文字列ブロックの先頭に終了タグがある場合および前記文字列ブロックの終端に開始タグがある場合にはこれを削除し、これ以外で前記文字列ブロック中に先行する開始タグに対応する後続する終了タグが存在する場合以外は対応する不足タグを追加することを特徴とする請求項３記載のハイパーメデイア記述言語による記述情報の抜粋表示システム。

【請求項６】　情報抽出手段はハイパーメデイア記述言語によることを示すタグと、情報のヘッダ部を示すタグと、情報の本文を示すタグとにより抜粋情報に整形することを特徴とする請求項３記載のハイパーメデイア記述言語による記述情報の抜粋表示システム。

【請求項７】　ハイパーメデイア記述言語による記述情報であるホームページ情報をネットワークを介して取得し前記ホームページ情報の指定されたキーワードを含む抜粋情報を作成して表示するすることを特徴とする請求項１記載のハイパーメデイア記述言語による記述情報の抜粋表示システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明はハイパーメデイア記述言語（以下ＨＴＭＬと略称する：ＨｙｐｅｒＴｅｘｔ

2

Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）による記述情報の抜粋表示システムに関し、特にWWW（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ）におけるＨＴＭＬで記述されたホームページ情報の中から与えられたキーワードに関連した部分を抜粋表示するＨＴＭＬ記述情報抜粋表示システムに関する。

【０００２】

【従来の技術】最近急速に普及してきたインターネットにおいては、WWWと呼ばれる情報検索システムが構築されており、このWWWによれば、種々の情報やサービスがＨＴＭＬと呼ばれる言語によって記述されたホームページにより提供されるようになされている。

【０００３】現在、ホームページの数は、莫大な数となっており、その中からユーザが自身によって所望するものを見つけ出すのは困難である。そこで、WWWでは、例えばキーワードなどを入力すると、そのキーワードを含むホームページを検索するような種々の検索サービスが提供されている。

【０００４】さらには特開平９－１７１５１３号公報記載の技術では、ユーザが、自身が所望するホームページが開設されたことを、容易に認識することもできるようになっている。

【０００５】そしてこのキーワード検索を行なって対応するホームページのＵＲＬ（Ｕｎｉｆｏｒｍ　Ｒｅｓｏｕｒｃｅ　Ｌｏｃａｔｅｒ、一種のアドレス）を得て、これを用いて、WWWブラウザによりホームページを表示してユーザは所望の情報を得ている。

【０００６】

【発明が解決しようとする課題】上述した従来のホームページの表示システム、すなわち、ＨＴＭＬ記述情報表示システムは、ホームページ、すなわち、ＨＴＭＬ記述情報の全文を表示しており、上記のキーワード検索で多数のＵＲＬが得られた場合にはその全てのホームページをWWWブラウザで表示して、本当に必要な情報の掲載されたページを目視で探さなければならないという欠点を有している。

【０００７】本発明の目的は、ホームページ、すなわち、ＨＴＭＬ記述情報の中から必要と考えられる部分を自動的に抜粋して一つに纏めて表示できるＨＴＭＬ記述情報抜粋表示システムを提供することにある。

【０００８】

【課題を解決するための手段】第１の発明のＨＴＭＬ記述情報抜粋表示システムは、ハイパーメデイア記述言語により記述された記述情報から指定されたキーワードを含むキーワード文字列を検索する文字列検索手段と、前記記述情報にあるタグを解析し予め指定された分割タグを検出するタグ解析手段と、前記検出された分割タグに基づいて前記キーワード文字列を前記記述情報から抽出しハイパーメデイア記述言語による抜粋情報に整形する情報抽出手段と、前記抽出されたキーワード文字列を表

示する表示手段とを含んで構成されている。

【０００９】第２の発明のＨＴＭＬ記述情報抜粋表示システムは、第１の発明のＨＴＭＬ記述情報抜粋表示システムにおいて、文字列検索手段は検索に際しては先ず記述情報からタグを削除してキーワード文字列を検索し検索後に前記削除したタグを復元することを特徴としている。

【００１０】第３の発明のＨＴＭＬ記述情報抜粋表示システムは、第１の発明のＨＴＭＬ記述情報抜粋表示システムにおいて、情報抽出手段はキーワードに先行する最も近い先行指定分割タグとキーワードに後続する最も近い後続指定分割タグとを検索して前記先行指定分割タグと後続指定分割タグとを含むその間の文字列を抽出することを特徴としている。

【００１１】第４の発明のＨＴＭＬ記述情報抜粋表示システムは、第３の発明のＨＴＭＬ記述情報抜粋表示システムにおいて、情報抽出手段は抽出したキーワードを含む文字列ブロックが複数あるときには文字列区切りタグにより１つの文字列ブロックにすることを特徴としている。

【００１２】第５の発明のＨＴＭＬ記述情報抜粋表示システムは、第３の発明のＨＴＭＬ記述情報抜粋表示システムにおいて、情報抽出手段は抽出した文字列ブロックの先頭に終了タグがある場合および前記文字列ブロックの終端に開始タグがある場合にはこれを削除し、これ以外で前記文字列ブロック中に先行する開始タグに対応する後続する終了タグが存在する場合以外は対応する不足タグを追加することを特徴としている。

【００１３】第６の発明のＨＴＭＬ記述情報抜粋表示システムは、第３の発明のＨＴＭＬ記述情報抜粋表示システムにおいて、情報抽出手段はハイパーメディア記述言語によることを示すタグと、情報のヘッダ部を示すタグと、情報の本文を示すタグとにより抜粋情報に整形することを特徴としている。

【００１４】第７の発明のＨＴＭＬ記述情報抜粋表示システムは、第１の発明のＨＴＭＬ記述情報抜粋表示システムにおいて、ハイパーメディア記述言語による記述情報であるホームページ情報をネットワークを介して取得し前記ホームページ情報の指定されたキーワードを含む抜粋情報を作成して表示するすることを特徴としている。

【００１５】

【発明の実施の形態】次に、本発明の実施の形態について図面を参照して説明する。

【００１６】図１は本発明のＨＴＭＬ記述情報抜粋表示システムの一実施の形態を示すブロック図である。

【００１７】本実施の形態のＨＴＭＬ記述情報抜粋表示システムは、図１に示すように、ＵＲＬとキーワードとを入力する入力部１１と、入力部１１から受け取り対応するＨＴＭＬファイルをＷＷＷサーバから取得するＨＴＭＬ取得部１２と、入力部１１からキーワードを受け取りＨＴＭＬ取得部１２からＨＴＭＬファイルを受けとりＨＴＭＬファイルに対しキーワード文字列の検索を行なう文字列検索部１３と、文字列検索部１３からキーワード文字列の検索の終了したＨＴＭＬファイルを受けとりそのタグを検索しキーワード文字周辺で文書構造の区切りに使用されることが多いタグを探し出すタグ解析部１４とキーワードを含みタグ解析部１４で探し出されたタグで囲まれた部分を抜き出し文書ブロックとしさらにその文書ブロックが複数ある場合はこれを一つにまとめ表示可能なＨＴＭＬファイル形式にする情報抽出部１５と、情報抽出部１５で得られるＨＴＭＬファイルを表示する出力部１６とから構成されている。

【００１８】これらにより構成される本実施の形態のＨＴＭＬ記述情報抜粋表示システムは、図２に示すキーボード等の入力装置２１と、情報を表示するデイスプレイ２２と、入力部１１、ＨＴＭＬ取得部１２、文字列検索部１３、タグ解析部１４、情報抽出部１５、出力部１６等の処理を行なうコンピュータ２３とにより実現される。

【００１９】図３は本実施の形態のＨＴＭＬ記述情報抜粋表示システムの動作を示す流れ図である。図１～図３を参照して本実施の形態のＨＴＭＬ記述情報抜粋表示システムの動作を説明する。

【００２０】まず、入力装置２１から与えられたＵＲＬとキーワードは、コンピュータ２３に実装された入力部１１によってＵＲＬはＨＴＭＬ取得部１２へ、キーワードは文字列検索部１３へ渡される（ステップ３０２）。

【００２１】ＵＲＬを受け取ったＨＴＭＬ取得部１２は対応するＨＴＭＬファイルをＷＷＷサーバから取得する（ステップ３０３）。

【００２２】次に、文字列検索部１３は入力部１１からキーワード、ＨＴＭＬ取得部１２からＨＴＭＬファイルを受け取り、ＨＴＭＬファイルに対しキーワード検索を行なう（ステップ３０４）。この際ＨＴＭＬのタグとコメント部分は検索を行なう前に一旦取り除き検索を行ない、検索終了後に元あった場所にタグとコメント部分を戻すようにした方がよい。これは、タグとコメントはＨＴＭＬファイルを参照する際、眼には触れない部分であり、キーワードの検索の対象にならないし、また、それがあることにより検索に誤作動を与えるためである。

【００２３】キーワードに当てはまる部分がなかった場合は（ステップ３０５のＮ枝）、その旨を出力部１６に伝え表示する（ステップ３１２）。

【００２４】キーワードに当てはまる部分があった場合は（ステップ３０５のＹ枝）、そのキーワードの位置を記憶して置く（ステップ３０６）。

【００２５】次にタグ解析部１４はキーワード検索の終了したＨＴＭＬファイルを受け取りその文書の区切りに使われるタグを検索する（ステップ３０７）。そして見

つけだされたタグを元にして情報抽出部１５はキーワードを含む部分を抜き出し文書ブロックにする（ステップ３０８）。文書ブロックの作成処理は図４を用いて後述する（ステップ３０８）。

【００２６】作成した文書ブロックが複数あるかチェックし（ステップ３０９）、複数あるときには（ステップ３０９のＹ枝）、改行タグ＜ＢＲ＞、水平線タグ＜ＨＲ＞等の区切り用タグを挟んでつなぎあわせ一つの文書ブロックにする（ステップ３１０）。この例は図５を用いて後述する。

【００２７】次に、ＨＴＭＬの必須タグ（図６（ｂ）に示す）を文書ブロックの前後に追加しＨＴＭＬファイルを作成する（ステップ３１１）。

【００２８】できたＨＴＭＬファイルは出力部１６からコンピュータ２３に表示される（ステップ３１２）。

【００２９】図４はタグを元にしてキーワードを含む文書ブロックをＨＴＭＬファイルから抜き出す処理を示す流れ図である。

【００３０】あらかじめ設定しておいた文書の構造的な区切りに使用されることの多いタグがＨＴＭＬファイル中のキーワードの前にあるかを検索し、キーワードの前にあって最も近い位置にあるタグを探す（ステップ４０２）。次に同様に設定したタグがキーワードより後ろにあるかを検索しキーワードの後ろにあって最も近い位置にあるタグを探す（ステップ４０３）。

【００３１】次に、キーワードの前後で見つかったタグを含む部分を抜き出す（ステップ４０４）。

【００３２】通常分割に利用するＨＴＭＬのタグは図６（ａ）に示すように、開始タグと終了タグが存在するが、抜き出した結果、開始、終了のいずれかが不足の場合と過剰の場合とが起るのでこれをチェックする（ステップ４０５）。文書ブロックの先頭に終了タグがある場合と、文書ブロックの終端に開始タグがある場合は、そのタグは過剰であるので取り去る。その他の場合で開始タグと終了タグが一致していない場合は、不足と見なしタグを追加する（ステップ４０６）。

【００３３】図５は文書ブロックをつなぎ併せる処理のステップ３１０の一例を示す流れ図である。文書ブロックが複数ある場合には、これらを一つの文書ブロックにまとめる。すなわち、ある文書ブロックＡにつづく次の文書ブロックＢがあるかを調べ（ステップ５０２）、なければ（ステップ５０２のＮ枝）、この動作は終了するが、あった場合には（ステップ５０２のＹ枝）、この文書ブロックＡの最後に＜ＨＲ＞タグを追加して（ステップ５０３）その後に次の文書ブロックＢを追加し（ステップ５０４）、残余の文書ブロックがなくなるまで＜ＨＲ＞タグで繋げてゆき、一つの文書ブロックに纏めてゆく。

【００３４】図７はＨＴＭＬファイルからキーワードを含む文書ブロックを図４に従って抜き出した結果の例を示している。

【００３５】以上説明したように、本実施の形態のＨＴＭＬ記述情報抜粋表示システムは、ホームページの中から必要と考えられる部分を自動的に抜粋して一つに纏めて表示することができ、本当に必要な情報の掲載されたページを従来よりも迅速にかつ、効率的に探索することができる。

【００３６】本実施の形態の説明ではＷＷＷのホームページを例にとって行なったが、本発明はこれに限定されるものではなく、ＨＴＭＬ記述言語による記述情報についての抜粋表示についても適用できることは自明である。

【００３７】

【発明の効果】以上説明したように、本発明のＨＴＭＬ記述情報抜粋表示システムは、ＨＴＭＬ記述言語による記述情報の中から必要と考えられる部分を自動的に抜粋して一つに纏めて表示することができ、本当に必要な情報の掲載されたページを従来よりも迅速にかつ、効率的に探索することができるという効果を有している。

【図面の簡単な説明】

【図１】本発明のＨＴＭＬ記述情報抜粋表示システムの一実施の形態を示すブロック図である。

【図２】本実施の形態のＨＴＭＬ記述情報抜粋表示システムにおける一実施例の構成を示すブロック図である。

【図３】本実施の形態のＨＴＭＬ記述情報抜粋表示システムの動作の一例を示す流れ図である。

【図４】本実施の形態のＨＴＭＬ記述情報抜粋表示システムの情報抽出の動作を示す詳細流れ図である。

【図５】本実施の形態のＨＴＭＬ記述情報抜粋表示システムの文書ブロックつなぎあわせ動作の詳細流れ図である。

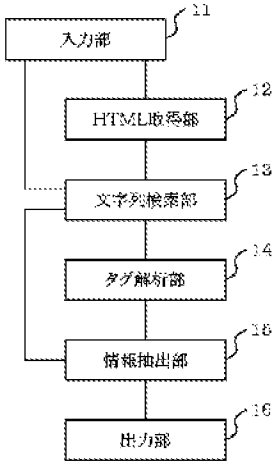
【図６】（ａ）は分割タグの一例を示すタグ図、（ｂ）は必須タグの一例を示すタグ図である。

【図７】ＨＴＭＬファイルからキーワードを含む文書ブロックを抜き出した結果の例を示す図である。

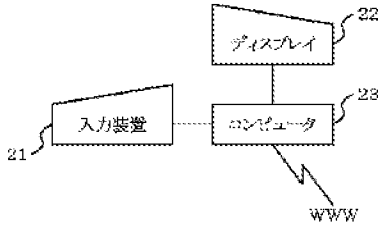
【符号の説明】

１１　入力部
１２　ＨＴＭＬ取得部
１３　文字列検索部
１４　タグ解析部
１５　情報抽出部
１６　出力部
２１　入力装置
２２　ディスプレイ
２３　コンピュータ

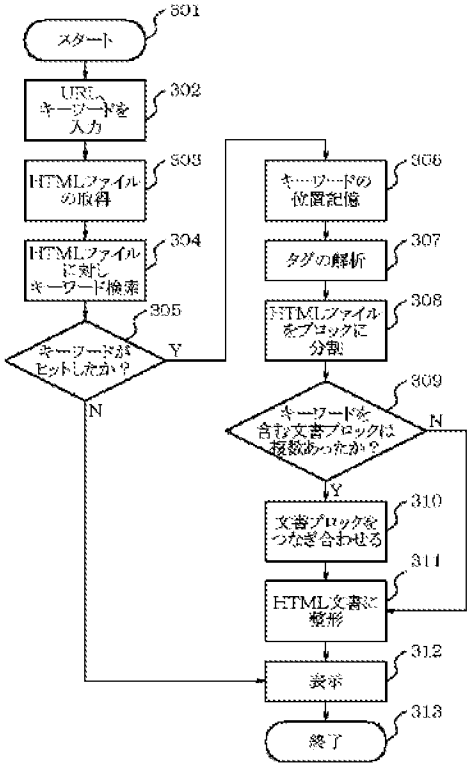
【図１】



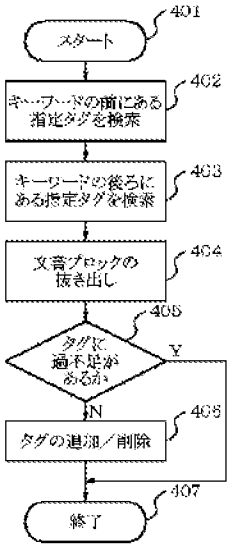
【図２】



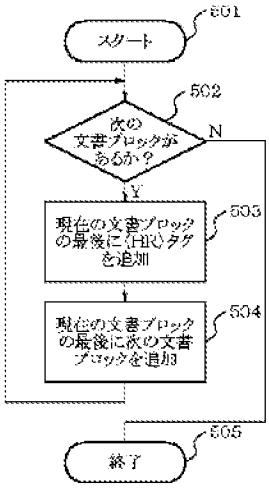
【図３】



【図４】



【図５】



【図６】

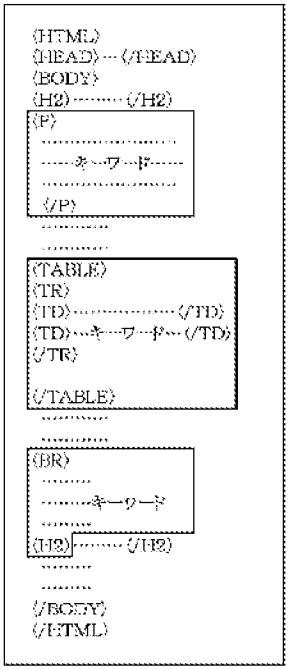
タグ		意味
開始	終了	
〈BODY〉	〈/BODY〉	書類の本文
〈H1〉	〈/H1〉	見出し
〈H2〉	〈/H2〉	見出し
〈H3〉	〈/H3〉	見出し
〈H4〉	〈/H4〉	見出し
〈H5〉	〈/H5〉	見出し
〈H6〉	〈/H6〉	見出し
〈TABLE〉	〈/TABLE〉	表の作成
〈P〉	〈/P〉	段落
〈CENTER〉	〈/CENTER〉	中央寄せ
〈OL〉	〈/OL〉	順番リスト
〈UL〉	〈/UL〉	非順番リスト
〈DL〉	〈/DL〉	定義リスト
〈PRE〉	〈/PRE〉	整形テキスト
〈BLOCKQUOTE〉	〈/BLOCKQUOTE〉	ブロックインデント
〈HR〉	--	水平線
〈BR〉	--	改行

(a)

タグ		意味
開始	終了	
〈HTML〉	〈/HTML〉	HTMLの最初と最後
〈HEAD〉	〈/HEAD〉	書類のヘッダ部
〈BODY〉	〈/BODY〉	書類の本文

(b)

【図7】


〈HTML〉
〈HEAD〉…〈/HEAD〉
〈BODY〉
〈H2〉……〈/H2〉
〈P〉
……
……キーワード……
……
〈/P〉
……
……
〈TABLE〉
〈TR〉
〈TD〉……〈/TD〉
〈TD〉…キーワード…〈/TD〉
〈/TR〉
〈/TABLE〉
……
……
〈BR〉
……
……キーワード
……
〈H2〉……〈/H2〉
……
……
〈/BODY〉
〈/HTML〉